

# Assignment 03

PUBH 8878

## Requirements:

- Show complete mathematical work for any derivations.
- Submit well-documented R code with clear comments and set seeds.
- Interpret results in a genetic/biological context where applicable.
- Submit the rendered PDF; do not submit the source .qmd.
- You may reuse and adapt your Lab 03 code, but your write-up must be self-contained.

## Problem 1: EM for ABO gene frequencies; derivation, inference, and sensitivity (25 pts)

### Part A (15 pts)

Consider phenotype counts from the ABO system under Hardy–Weinberg equilibrium (HWE):  $n_A, n_{AB}, n_B, n_O$  with total  $N$ . Let genotype frequencies be  $p_A^2, 2p_Ap_O, 2p_Ap_B, p_B^2, 2p_Bp_O, p_O^2$  and  $p_O = 1 - p_A - p_B$ .

Derive the EM updates shown in lecture. Specifically, show that the E-step allocations for  $A$  and  $B$  phenotypes are

$$\tilde{n}_{AA} = n_A \frac{p_A^2}{p_A^2 + 2p_Ap_O}$$

$$\tilde{n}_{AO} = n_A \frac{2p_Ap_O}{p_A^2 + 2p_Ap_O}$$

and similarly for  $BB, BO$ , and that the M-step is

$$p_A^{(t+1)} = \frac{2\tilde{n}_{AA} + \tilde{n}_{AO} + n_{AB}}{2N}$$

$$p_B^{(t+1)} = \frac{2\tilde{n}_{BB} + \tilde{n}_{BO} + n_{AB}}{2N}$$

$$p_O^{(t+1)} = 1 - p_A^{(t+1)} - p_B^{(t+1)}$$

Hint (how to derive EM generally):

- 1) Choose latent variables so the complete data are simple. Here, treat latent genotype counts,

$$(n_{AA}, n_{AO}, n_{AB}, n_{BB}, n_{BO}, n_{OO})$$

as missing, with only phenotype totals observed.

- 2) Write the complete-data log-likelihood

$$\ell_c(p_A, p_B) = n_{AA} \log p_A^2 + n_{AO} \log(2p_A p_O) + n_{AB} \log(2p_A p_B) +$$

$$n_{BB} \log p_B^2 + n_{BO} \log(2p_B p_O) + n_{OO} \log p_O^2$$

using  $p_O = 1 - p_A - p_B$ .

3. E-step: replace latent counts by their conditional expectations given the observed phenotypes under current parameters, e.g., for phenotype  $A$ , and form  $Q(p | p^{(t)}) = \mathbb{E}[\ell_c(p) | \text{data}, p^{(t)}]$ .
4. M-step: M-step: Maximize  $Q(p | p^{(t)})$  subject to  $p_A + p_B + p_O = 1$ . Hint: Consider rewriting the objective in terms of allele counts rather than genotype counts.

## Part B (10 pts)

Consider two biallelic SNPs with haplotypes  $\{ab, aB, Ab, AB\}$  under HWE and unphased genotypes  $(g_1, g_2) \in \{0, 1, 2\}^2$ . Note that  $(ab, ab)$  yields  $(0, 0)$ ,  $(ab, aB)$  or  $(ab, Ab)$  yields  $(0, 1)$ ,  $(AB, AB)$  yields  $(2, 2)$ , etc.

Show that if every observed genotype is  $(1, 1)$ , then  $P\{(1, 1)\} = 2(p_{ab}p_{AB} + p_{aB}p_{Ab})$  and the likelihood depends only on the cross-sum  $S = p_{ab}p_{AB} + p_{aB}p_{Ab}$  (a ridge; parameters not identifiable).

## Problem 2: Two-point linkage — LOD and support intervals (25 pts)

Suppose one heterozygous transmitting parent ( $A/a$  and  $B/b$ ) is crossed to an  $aabb$  mate, yielding child haplotype counts  $(n_{AB}, n_{Ab}, n_{aB}, n_{ab})$ . Let  $n_{NR} = n_{AB} + n_{ab}$  and  $n_R = n_{Ab} + n_{aB}$ .

**Part A: LOD from counts (10 pts).**

Show that for a given recombination fraction  $\theta \in (0, 0.5)$  the two-point LOD relative to independence ( $\theta = 0.5$ ) is

$$\text{LOD}(\theta) = \log_{10} \left\{ \frac{\theta^{n_R} (1 - \theta)^{n_{NR}}}{0.5^{n_R + n_{NR}}} \right\}.$$

Derive the MLE  $\hat{\theta}$  and show it equals  $n_R / (n_R + n_{NR})$  when  $0 < \hat{\theta} < 0.5$ .

**Part B: Unknown phase and LD-informed LOD (15 pts)**

A heterozygous transmitting parent ( $A/a$ ,  $B/b$ ) has **unknown phase**: either **coupling** ( $AB/ab$ ) or **repulsion** ( $Ab/aB$ ). Let

$$n_{NR} = n_{AB} + n_{ab}, \quad n_R = n_{Ab} + n_{aB}, \quad N = n_{NR} + n_R.$$

Let  $w = \Pr\{\text{coupling } (AB/ab)\}$  and  $1 - w = \Pr\{\text{repulsion } (Ab/aB)\}$ .

**(i) Mixture likelihood (5 pts).**

Show that with unknown phase the observed-data likelihood is a **mixture** of the two phase-specific binomial likelihoods:

$$L(\theta; w) = w (1 - \theta)^{n_{NR}} \theta^{n_R} + (1 - w) (1 - \theta)^{n_R} \theta^{n_{NR}}.$$

Hence the two-point LOD relative to independence ( $\theta = 0.5$ ) is

$$\text{LOD}(\theta; w) = \log_{10} \left\{ \frac{w (1 - \theta)^{n_{NR}} \theta^{n_R} + (1 - w) (1 - \theta)^{n_R} \theta^{n_{NR}}}{0.5^N} \right\}.$$

**(ii) Linking LD to  $w$  (5 pts).**

Let population haplotype frequencies be  $p = (p_{ab}, p_{aB}, p_{Ab}, p_{AB})$  (sum to 1).

Condition on the parent being the **double heterozygote** ( $g_1, g_2$ ) = (1, 1). Use Bayes' rule to show

$$w = \Pr\{(ab, AB) \mid (1, 1)\} = \frac{p_{ab} p_{AB}}{p_{ab} p_{AB} + p_{aB} p_{Ab}},$$

$$1 - w = \frac{p_{aB} p_{Ab}}{p_{ab} p_{AB} + p_{aB} p_{Ab}}.$$

Define  $D = p_{ab} p_{AB} - p_{aB} p_{Ab}$  and note that  $\text{sign}(D)$  indicates whether **coupling** ( $D > 0$ ) or **repulsion** ( $D < 0$ ) phase is a priori more likely.

**(iii) Quick numerical check (5 pts).**

Take  $p^* = (0.40, 0.10, 0.25, 0.25)$ .

Compute  $w$  and  $1 - w$ . Then, using the example counts  $(n_{AB}, n_{Ab}, n_{aB}, n_{ab}) = (18, 5, 4, 17)$  (so  $N = 44$ ,  $n_R = 9$ ,  $n_{NR} = 35$ ), evaluate and **compare**  $\text{LOD}(\hat{\theta}; w = \frac{1}{2})$  versus  $\text{LOD}(\hat{\theta}; w)$  at  $\hat{\theta} = n_R/N$ .

Briefly explain (one sentence) how LD information ( $w \neq \frac{1}{2}$ ) can increase or decrease the peak LOD when phase is unknown.

**Problem 3: Two-SNP haplotype EM and LD measures (25 pts)**

**Part A (10 pts)**

For two biallelic SNPs with haplotypes  $\{ab, aB, Ab, AB\}$  at frequencies  $\{p_{ab}, p_{aB}, p_{Ab}, p_{AB}\}$  (summing to 1), enumerate the possible haplotype pairs consistent with each unphased genotype  $(g_1, g_2) \in \{0, 1, 2\}^2$ .

Show that only  $(g_1, g_2) = (1, 1)$  is ambiguous with two possible pairs:  $(ab, AB)$  and  $(aB, Ab)$ .

**Part B (10 pts)**

Simulate  $N = 1000$  individuals from true haplotype frequencies  $p^* = (0.40, 0.10, 0.25, 0.25)$  and estimate  $\hat{p}$  via EM from a uniform start. Report  $\hat{p}$  and absolute errors  $|\hat{p} - p^*|$ .

Note that the E-step weights for  $(1, 1)$  are proportional to  $p_{ab}p_{AB}$  and  $p_{aB}p_{Ab}$ , and the M-step update is  $p^{(t+1)} = (\text{expected hap counts})/(2N)$ .

Here is a sample R code snippet to get you started:

```
set.seed(8878)

N <- 1000
p_true <- c(ab = 0.40, aB = 0.10, Ab = 0.25, AB = 0.25)

# helper: draw N unordered haplotype pairs, then make unphased genotypes (g1,g2)
draw_genotypes <- function(N, p) {
  # code haplotypes to allele counts (B allele) at SNP1, SNP2
  H <- rbind(ab = c(0, 0), aB = c(0, 1), Ab = c(1, 0), AB = c(1, 1))
  hap1 <- sample(rownames(H), size = N, replace = TRUE, prob = p)
  hap2 <- sample(rownames(H), size = N, replace = TRUE, prob = p)
  G <- H[hap1, ] + H[hap2, ] # N x 2 matrix with entries in {0,1,2}
  as.data.frame(G) |>
    setNames(c("g1", "g2"))
}
```

```

}

# simulate data
dat <- draw_genotypes(N, p_true)

```

### Part C (5 pts)

Compute  $D = p_{11} - p_{B1}p_{B2}$  with  $p_{11} = p_{AB}$ ,  $p_{B1} = p_{Ab} + p_{AB}$ ,  $p_{B2} = p_{aB} + p_{AB}$ . Report  $D$  and  $r^2 = D^2 / (p_{B1}(1 - p_{B1})p_{B2}(1 - p_{B2}))$ . Comment on how LD would affect single-marker association at either SNP.

### Problem 4: Single-marker association with QC and LD attenuation (25 pts)

Simulate  $n = 2000$  unrelated individuals. Let a causal biallelic SNP  $C$  have MAF 0.30 and generate a quantitative trait  $Y$  with additive effect size  $\beta_C = 0.50$  (per allele) and noise  $\epsilon \sim \mathcal{N}(0, 1)$ . Let a tag SNP  $T$  be in LD with  $C$  such that  $r = \text{corr}(G_C, G_T) = 0.8$  and both are in HWE. Let  $T$  have MAF 0.30.

#### Part A (15 pts)

Generate  $(G_C, G_T, Y)$  by first simulating haplotypes for  $(C, T)$  with a chosen LD structure that yields  $r \approx 0.8$ , then form genotypes and  $Y = \beta_C G_C + \epsilon$ . Fit simple linear models  $Y \sim G_C$  and  $Y \sim G_T$  and report  $\hat{\beta}_C$  and  $\hat{\beta}_T$ , alongside their 95% confidence intervals.

#### Part B (10 pts)

Introduce a basic QC step: test HWE in the controls of a case-control subsample formed by thresholding  $Y$  at its 80th percentile to define cases (cases are the top 20% of  $Y$ ). Compute an exact or  $\chi^2$  HWE p-value in controls for  $T$ ; state whether you would flag  $T$  using a threshold of  $10^{-6}$  and why QC is typically done in controls only.